Environmental Health Perspectives
Vol. 32, pp. 113-123, 1979

Review of the Methods of Quantification

by Yutaka Tanaka*

In order to analyze qualitative observations, methods of quantification or optimal scaling have been proposed by Fisher, Guttman, and Hayashi. According to these methods, scores are assigned optimally in some objective and operational sense to the qualitative categories.

The present paper mainly reviews Hayashi's four methods of quantification from the mathematical point of view. They are widely used, especially in Japan, in various fields such as social and marketing surveys, psychological research, medical research, etc., where information is obtained mainly in the form of qualitative categories. The first and second methods are applied to the case where an external criterion is present, and are used to predict the external criterion or to analyze the effects of factors. On the other hand, the third and fourth methods are applied to the case where no external criterion is present, and are used to construct a spatial configuration so as to grasp the mutual relationship of the data.

After reviewing Hayashi's four methods, we discuss two topics which have been pointed out as the problems to be solved in applying the methods of quantification. One is quantification for ordered categories and the other is statistical consideration. With respect to these topics we review some recently developed methods including the studies due to the present author. Finally we mention briefly several computer programs available in Japan.

Introduction

In some experimental and observational studies, the responses and/or attributes of subjects are measured only by qualitative categories. In order to analyze such observations, methods of quantification or optimal scaling have been proposed by, among others, Fisher (*1*), Guttman (*2*), and Hayashi (*3-9*). According to these methods, scores are assigned optimally in some objective and operational sense to these qualitative categories. In Japan, Hayashi's methods of quantification are well known and widely used in various fields, such as social and marketing surveys, psychological research, and medical research, where information is obtained mainly in the form of qualitative categories.

The main purpose of the present paper is to review Hayashi's four methods of quantification. They are explained mainly from the mathematical point of view. Then, in addition, we focus on two topics, which have been pointed out as the problems to be solved in using the methods of quantification: the methods of quantification for ordered categories and the statistical considerations. Finally we mention briefly several computer programs available in Japan.

*Department of Statistics, The School of General Education, Okayama University, 2-1-1 Tsushima, Okayama 700, Japan.

Hayashi's Four Methods of Quantification

Among various methods proposed by Hayashi (*3-9*), especially the four methods shown in Table 1 are widely applied in Japan and called simply as Hayashi's first-fourth methods of quantification. As shown in Table 1, they are divided into two main classes. One contains the methods for the case where an external criterion is present and is used to predict the external criterion or to analyze the effects of factors. The other contains the methods for the case where no external criterion is present, and is used to construct a spatial configuration so as to grasp the mutual relationships of the data. The "external criterion", which is also called "outside variable", means something to be predicted or explained.

First Method of Quantification (Quantification I)

The first method of quantification is a method to predict the quantitative external criterion or criter-

Table 1. Hayashi's four methods of quantification.

Situation	Observation	Method
Case with an external criterion (for prediction or analyzing the effects of factors)	The external criterion is observed quantitatively	First method (to maximize the correlation coefficient)
	The external criterion is observed qualitatively	Second method (to maximize the correlation ratio)
Case with no external criterion (for classification or constructing a spatial configuration)	Response patterns of subjects on some attributes are given	Third method (to maximize the correlation coefficient between subjects and categories)
	Similarities between pairs of subjects are observed quantitatively	Fourth method (to maximize the objective function [Eq. (53)]

Table 2. Data for the first method of quantification.

External criterion	Item 1				Item 2				...	Item I			
	1	2	...	c_1	1	2	...	c_2	...	1	2	...	c_I
Y_1	√					√						-	√
Y_2		√			√					√			
.													
.													
.													
Y_n				√				√			√		

ion variable on the basis of the information concerning the qualitative attributes of each subject and to analyze the influence of each attribute to the criterion variable. The data for this method are usually given in the form of Table 2.

Let Y be the quantitative external criterion, and let us suppose that every subject under study can be classified into one and only one of c_i categories of the i-th attribute item for $i = 1, 2, \ldots, I$. Dummy variables are introduced such that

$$x_\alpha(ij) = \begin{cases} 1, \text{if subject } \alpha \text{ belongs to category } j \text{ of the} \\ \quad i\text{-th attribute item}, i = 1, 2, \ldots, I \\ 0, \text{otherwise}, \end{cases} \quad (1)$$

for subject α, $\alpha = 1, 2, \ldots, n$. In order to analyze the relationships between the external criterion and the qualitative attributes we shall assign a quantity or numerical score s_{ij} to category j of the i-th item, and as a result assign a score

$$W_\alpha(i) = \sum_{j=1}^{c_i} s_{ij} x_\alpha(ij) \qquad i = 1, 2, \ldots, I \quad (2)$$

to attribute i of subject α and a score

$$Y_{(c)\alpha} = W_\alpha(1) + w_\alpha(2) + \ldots + W_\alpha(I) = \sum_{i=1}^{I} \sum_{j=1}^{c(i)} s_{ij} x_\alpha(ij)$$

where

$$c_j \equiv c(j) \quad (3)$$

to subject α. The principle for quantification is to maximize the sample correlation coefficient between $\{Y_\alpha\}$ and $\{Y_{(c)\alpha}\}$, i.e.,

$$\rho^2 = r^2(Y, Y_{(c)}) = \frac{\{\Sigma_\alpha (Y_\alpha - \bar{Y})(Y_{(c)} - \bar{Y}_{(c)})\}^2}{\Sigma_\alpha (Y_\alpha - \bar{Y})^2 \Sigma_\alpha (Y_{(c)\alpha} - \bar{Y}_{(c)})^2} \to \max. \quad (4)$$

The basic idea is to predict the external criterion as accurately as possible on the basis of a linear combination, Eq. (3). Due to the fact that the principle of maximizing the correlation coefficient is equivalent to that of minimizing the mean-square error (*10*), we obtain the normal equation (5) from the theory of ordinary linear regression:

$$\sum_j \sum_k f(uv, jk)\, s_{jk} = \Sigma_\alpha Y_\alpha x_\alpha(uv) \qquad u = 1, 2, \ldots, I, \quad v = 1, 2, \ldots, c_u \quad (5)$$

where

$$f(uv, jk) = \Sigma_\alpha x_\alpha(jk)\, x_\alpha(uv) \quad (6)$$

The optimal scores for the categories of the attributes are obtained by solving the linear simultaneous equation (5). Then, using the optimal scores $\{s_{ij}\}$, each qualitative attribute is quantified by Eq. (2) and the external criterion Y can be predicted by Eq. (3).

It may be considered that the efficiency of quantification is high when the multiple correlation coefficient or $R^2 = \rho_{\max}{}^2$ is large, but it is low when R^2 is small. The contribution of the i-th attribute to the external criterion is measured by the partial correlation coefficient

$$r\,[Y \cdot W(i);\, W(1),\, \ldots,\, W(i-1),\, W(i+1),\, \ldots,\, W(I)] \tag{7}$$

or approximately, by the range of the assigned scores

$$R_i = \max_j S_{ij} \min_j S_{ij} \tag{8}$$

In actual data analysis, partial correlation coefficients and/or ranges are often represented graphically for the convenience to find important attributes or factors.

No probabilistic model is assumed in the first method of quantification. However, if the problem is recognized as the multiple regression of the external criterion $\{Y_\alpha\}$ on the dummy variables $\{x_\alpha(ij)\}$ such that

$$Y_\alpha = \sum_i \sum_j \theta_{ij}\, x_\alpha(ij) + e_\alpha \alpha \qquad \alpha = 1, 2, \ldots, m \tag{9}$$

and if the normality of the error term e_α can be assumed, the statistical properties of the scores or estimates s_{ij} for θ_{ij} are derived by the theory of regression, and the contribution of each attribute can be tested exactly by using the ordinary significance test of regression coefficients.

Second Method of Quantification (Quantification II)

The second method of quantification is a method to predict the qualitative external criterion on the basis of the information concerning the qualitative attributes of each subject and to analyze the influence of each attribute to the discrimination of the external criterion. The data for this method are usually given in the form of Table 3. It is formulated in the following two ways.

Formulation Based on Canonical Analysis. We suppose there exists an external criterion with r categories or groups $\pi_1, \pi_2, \ldots, \pi_r$, and introduce the following dummy variables:

$$x_{ij}(kl) = \begin{cases} 1, & \text{if the } j\text{-th subject in } \pi_i \text{ belongs to} \\ & \text{category } l \text{ of the } k\text{-th attribute} \\ 0, & \text{otherwise,} \end{cases} \quad k = 1, 2, \ldots, I \tag{10}$$

In order to analyze the relationships between the external criterion and the qualitative attributes we shall assign a numerical score s_{kl} to category l of the k-th item as in the case of the first method of quantification, and as a result assign a score

$$W_{ij}(k) = \sum_{l=1}^{c(k)} s_{kl}\, x_{ij}(kl) \tag{11}$$

to qualitative attribute k of the j-th subject in π_i, and a score

$$\begin{aligned} Y_{(c)ij} &= W_{ij}(1) + W_{ij}(2) + \ldots + W_{ij}(I) \\ &= \sum_k \sum_l s_{kl}\, X_{ij}(kl) \end{aligned} \tag{12}$$

to the j-th subject in π_i. The principle of quantification is to maximize the sample correlation ratio or the between-groups variation relative to the total variation, i.e.,

$$\eta^2 = S_B/S_T \to \max. \tag{13}$$

where

$$\begin{aligned} S_B &= \sum_i \sum_j (\bar{Y}_{(c)i\cdot} - \bar{Y}_{(c)\cdot\cdot})^2 \\ &= \sum_k \sum_l \sum_u \sum_v \left\{ \sum_i \frac{g^i(kl)\, g^i(uv)}{n_i} - \frac{n'_{kl} n'_{uv}}{n} \right\} s_{kl} s_{uv} \end{aligned} \tag{14}$$

$$\begin{aligned} S_T &= \sum_i \sum_j (Y_{(c)ij} - \bar{Y}_{(c)\cdot\cdot})^2 \\ &= \sum_k \sum_l \sum_u \sum_v \left\{ f(kl, uv) - \frac{n'_{kl} n'_{uv}}{n} \right\} s_{kl} s_{uv} \end{aligned} \tag{15}$$

Using the matrix notations such that

$$\mathbf{S} = [s_{11}, s_{12}, \ldots, s_{1c(1)}, \ldots, s_{I1}, \ldots, s_{Ic(I)}]' : \sum_i c_i \times 1$$

$$\mathbf{B} = [b(uv, kl)] : \sum_i e_i \times \sum_i c_i$$

$$\mathbf{T} = [f(uv,kl) - n'_{uv} n'_{kl}/n] : \sum_i c_i \times \sum_i c_i$$

$$b(uv,kl) = \sum_i \frac{g^i(kl) g^i(uv)}{n_i} - \frac{n'_{kl} n'_{uv}}{n}$$

n'_{uv} = the number of subjects belonging to category v of the u-th item,

$g^i(uv)$ = the number of subjects belonging to category v of the u-th item in the i-th group,

$f(uv, kl)$ = the number of subjects belonging to category v of the u-th item and category l of the k-th item simultaneously,

the optimization problem (13) is expressed as the problem to maximize the ratio of quadratic forms, i.e.,

$$\eta^2 = \mathbf{s'Bs}/\mathbf{s'Ts} \to \max. \tag{16}$$

and is transformed to the eigenvalue problem

Table 3. Data for the second method of quantification.

External criterion		Item 1				Item 2				...	Item I			
		1	2	...	c_1	1	2	...	c_2	...	1	2	...	c_I
1	1		√			√								√
	2	√							√			√		
	.													
	.													
	.													
	n_1				√			√			√			
.														
.														
.														
r	1	√				√					√			
	2				√				√				√	
	.													
	.													
	.													
	n_r		√			√								√

$$(\mathbf{B} - \eta^2\mathbf{T})\,\mathbf{s} = \mathbf{0} \tag{17}$$

Due to the condition of exclusive and exhaustive categories there exist linear dependencies among the dummy variables such that

$$\sum_l x_{ij}(kl) = 1,$$
$$k = 1, 2, \ldots, \mathrm{I}, \text{ for any } i,j \tag{18}$$

Then without any loss of generality we may exclude each dummy variable for any arbitrary category per item and the corresponding rows and columns of the matrices **B** and **T**. It is just the same to assign zero scores to such categories. After solving

$$(\tilde{\mathbf{B}} - \eta^2\tilde{\mathbf{T}})\,\tilde{\mathbf{s}} = \mathbf{0} \tag{19}$$

where the matrices with the tilde indicate the abbreviated matrices, we may normalize the location to satisfy the relation

$$\sum_l n'_{kl}\, s_{kl} = 0$$
$$k = 1, 2, I \tag{20}$$

if necessary. The number of nonzero eigenvalues is generally given by min $[r - 1, \sum_i (c_i - 1)]$ except for degenerated cases.

The optimization problem (16) is interpreted as the application of canonical analysis for dummy variables $\{x_{ij}\,(kl)\}$.

Furthermore, the scores assigned to the categories of the external criterion are defined by the mean values of $Y_{(c)ij}$ within the groups, i.e.,

$$Y_{(c)i.} = \frac{1}{n_i}\sum_j Y_{(c)ij} = \sum_j \sum_k \sum_l s_{kl} x_{ij}(kl)$$
$$i = 1, 2, \ldots, r \tag{21}$$

where n_i denotes the sample size of π_i.

Formulation Based on Canonical Correlation Analysis. In the above formulation the dummy variables were introduced for the categories of qualitative attributes. Now we shall introduce the dummy variables for the categories of not only the qualitative attributes but also the qualitative external criterion, i.e.

$$z_\alpha(i) = \begin{cases} 1, & \text{if subject } \alpha \text{ belongs to category } i \\ & \text{of the external criterion} \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

$$x_\alpha(kl) = \begin{cases} 1, & \text{if subject } \alpha \text{ belongs to category } l \\ & \text{of the } k\text{-th attribute item} \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

In order to analyze the relationships between the qualitative external criterion and the qualitative attributes we shall assign numerical scores s_{kl} to category l of the k-th attribute item and t_i to category i of the external criterion.
Then the quantities

$$W_{(c)\alpha} = \sum_{i=1}^{r} t_i z_\alpha(i) \tag{24}$$

and

$$Y_{(c)\alpha} = \sum_{k=1}^{I} \sum_{l=1}^{c_k} s_{kl} X_\alpha(kl) \tag{25}$$

are given to subject α from the viewpoints of the qualitative external criterion and the qualitative attributes, respectively. Now we shall introduce the principle of quantification to maximize the sample correlation coefficients between $\{W_{(c)\alpha}\}$ and $\{Y_{(c)\alpha}\}$, or in other words, the sample canonical correlation coefficient between the two sets of dummy variables

$\{z_\alpha(i), i = 1, 2, \ldots, r\}$ and $\{x_\alpha(kl), k = 1, 2, \ldots, I, l = 1, 2, \ldots, c_k\}$, i.e.,

$$r^2(W_{(c)}, Y_{(c)}) \rightarrow \max. \tag{26}$$

Let us use the matrix notations such that

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}$$

where S is the sample variance-covariance matrix of the dummy variables $\{z_\alpha(i)\}$ and $\{x_\alpha(kl)\}$,

$\mathbf{S}_{11}$ = The $r \times r$ matrix with $[n_i\delta_{ij} - (n_in_j/n)]/n$ as its (i,j) element

$\mathbf{S}_{12}$ = the $r \times \sum_k c_k$ matrix with $[g^i(kl) - (n_in'_{kl})/n]/n$ as its (i,kl) element

$\mathbf{S}_{22}$ = the $\sum_k c_k \times \sum_k c_k$ matrix with $[f(kl\ uv) - (n'_{kl}n'_{uv})/n]/n$ as its (kl, uv) element,

$\mathbf{t} = [t_1, t_2, \ldots, t_r]'$

$\mathbf{s} = [s_{11}, \ldots, s_{ic(i)}, \ldots, s_{Ii}, \ldots, s_{Ic(I)}]'$

Then the above principle is expressed as

$$\rho^2 = r^2(W_{(c)}, Y_{(c)}) = \frac{(\mathbf{t}\,\mathbf{S}_{12}\,\mathbf{s})^2}{(\mathbf{t}'\,\mathbf{S}_{11}\mathbf{t})(\mathbf{s}'\,\mathbf{S}_{22}\mathbf{s})} \rightarrow \max. \tag{27}$$

As noted previously there exist linear dependencies among the dummy variables, we may exclude each dummy variable for an arbitrary category per item and the corresponding rows and columns of the matrix $\mathbf{S}_{ij}$, $i, j = 1, 2$. Denoting such abbreviated matrices with the superimposed $(\sim)$, we obtain

$$\rho^2 = \frac{(\tilde{\mathbf{t}}'\,\tilde{\mathbf{S}}_{12}\,\tilde{\mathbf{s}})'}{(\tilde{\mathbf{t}}'\,\tilde{\mathbf{s}}_\pi\,\tilde{\mathbf{t}})\,(\tilde{\mathbf{s}}'\,\tilde{\mathbf{S}}_{22}\,\tilde{\mathbf{s}})} \rightarrow \max. \tag{28}$$

Hence, due to the theory of canonical correlation analysis, the optimal scores satisfying (28) are given by solving the simultaneous equations such as

$$\begin{aligned} -\rho\,\tilde{\mathbf{S}}_{11}\tilde{\mathbf{t}} + \tilde{\mathbf{S}}_{12}\,\tilde{\mathbf{s}} &= \mathbf{0} \\ \tilde{\mathbf{S}}_{21}\,\tilde{\mathbf{t}} - \rho\,\tilde{\mathbf{S}}_{22}\,\tilde{\mathbf{s}} &= \mathbf{0} \end{aligned} \tag{29}$$

which are transformed into the following two types of eigenvalue problems with the common eigenvalues.

$$(\tilde{\mathbf{S}}_{12}\,\tilde{\mathbf{S}}_{22}^{-1}\,\tilde{\mathbf{S}}_{21} - \rho^2\,\tilde{\mathbf{S}}_{11})\,\tilde{\mathbf{t}} = 0 \tag{30}$$

$$(\tilde{\mathbf{S}}_{21}\,\tilde{\mathbf{S}}_{11}^{-1}\,\tilde{\mathbf{S}}_{12} - \rho^2\,\tilde{\mathbf{S}}_{22})\,\tilde{\mathbf{s}} = 0 \tag{31}$$

The optimal scores for the categories of the external criterion and the attributes are given by the eigenvector corresponding to the maximum eigenvalue.

Concerning the relationship between the results, the following are derived.

Since the inverse of the matrix

$$\tilde{\mathbf{S}}_{11} = (1/n)\ \mathrm{diag}\,[n_2, \ldots, n_r] + [n_2/n, \ldots, n_r/n]\,[n_2/n, \ldots, n_r/n]' \tag{32}$$

can be expressed explicitly by

$$\tilde{\mathbf{S}}_{11}^{-1} = n\ \mathrm{diag}\,[1/n_2, \ldots, 1/n_r] + (n/n_1)\,[1, \ldots, 1]\,[1, \ldots, 1]' \tag{33}$$

the matrices in the eigenvalue problem (31) are obtained as the (kl, uv) element of

$$\tilde{\mathbf{S}}_{21}\,\tilde{\mathbf{S}}_{11}^{-1}\,\tilde{\mathbf{S}}_{12} = \frac{1}{n}\left\{\sum_i \frac{g^i(kl)\,g^i(uv)}{n_i} - \frac{n'_{kl}n'_{uv}}{n}\right\} \tag{34}$$

and the (kl, uv) element of

$$\tilde{\mathbf{S}}_{22} = \frac{1}{n}\left\{f(kl, uv) - \frac{n_{kl}'n_{uv}'}{n}\right\} \tag{35}$$

Hence

$$\begin{aligned} n\,\tilde{\mathbf{S}}_{21}\,\tilde{\mathbf{S}}_{11}^{-1}\,\tilde{\mathbf{S}}_{12} &= \mathbf{B} \\ n\,\tilde{\mathbf{S}}_{22} &= \tilde{\mathbf{T}} \end{aligned} \tag{36}$$

Thus the two eigenvalue problems (19) and (31) are equivalent. Furthermore, it becomes clear that the relationships

$$t_i = (\bar{Y}_{(c)i.} + \text{const.})/\rho \qquad i = 1, 2, \ldots, r \tag{37}$$

hold about the scores for categories of the external criterion and that they are equivalent to each other except for the normalization of location and scale. However, the formulation based on canonical correlation analysis is more appropriate in view of the quantification of the external criterion, and more convenient to treat ordered categories of the external criterion or to derive the asymptotic properties of the sample optimal scores.

Using the optimal scores $\{s_{kl}\}$ and $\{t_i\}$, the qualitative attributes and the qualitative external criterion are quantified by Eqs. (25) and (24). It may be said, as in the case of the first method of quantification, that the efficiency of quantification is high when the correlation ratio η^2 (or correlation coefficient ρ^2) is large, but it is low when η^2 is small. The contribution of the i-th attribute to the external criterion is measured by the partial correlation coefficient

$$r[W_{(c)} \cdot W(i); W(1), \ldots, W(i-1), W(i+1), \ldots, W(I)] \tag{38}$$

or approximately by the range of the assigned scores

$$R_i = \max_j s_{ij} - \min_j s_{ij} \tag{39}$$

When the discrimination among the categories of the external criterion is not satisfactory by assigning unidimensional scores, we may use multidimensional scores. In such cases the eigenvectors corresponding to the eigenvalues smaller than the largest

should be used. The principle becomes the maximization of $\Pi_i \eta_i^2$ instead of η^2 under the orthogonality constraints,

$$\mathbf{s}_i' \mathbf{T} \mathbf{s}_j = 0 \quad \text{for } i \neq j \qquad (40)$$

Hayashi (*6*) discussed precisely the multidimensional case.

Fisher (*1*) proposed a method to quantify the response categories by the principle to maximize the variation due to the effects of factors relative to the total variation in a two-way analysis of variance. It gives the same result with that of Hayashi's second method when the response is chosen as the external criterion. A similar principle was also applied by Johnson (*11*).

For the investigation of a factor-response relationship, Hayashi's second method may be applied in the following two different manners. One is the case where a response item is chosen as the external criterion and the problem is to predict the response from the qualitative factors by a similar way as the regression analysis. It just corresponds to Fisher's method. The other is the case where a factor is chosen as the external criterion and the problem is to discriminate between groups corresponding to the categories of the chosen factor by a similar way as the canonical analysis. In relation to these two situations several generalized principles were proposed to quantify a single or multiple responses on the basis of a univariate or multivariate linear model by Tanaka and Asano (*12, 13*) and Tanaka (*14*).

Third Method of Quantification (Quantification III)

Suppose that response patterns to categories of qualitative attributes are given in the form of Table 4. In this table the subjects showing a same response pattern are pooled into one row, and the frequency of each response pattern is denoted by s_i, $i = 1, 2, \ldots, Q$. The basic idea of the third method of quantification is to arrange the rows and columns so that those which resemble to each other are gathered together.

Table 4. Data for the third method of quantification.

	Attribute category							
Subject	1	2	3	...	j	... —	R	Frequency
1	✓		✓				✓	f_1
2		✓					✓	f_2
3								f_3
.								
.								
.								
j		✓			✓			f_j
.								
.								
.								
Q			✓				✓	f_Q

Now let us assign numerical scores y_i and x_j to subject i and category j, respectively. Then the event that subject i responds to category j is expressed by a pair of the numerical scores (y_i, x_j). The above basic idea corresponds to the principle to maximize the correlation coefficient between x and y. We define a dummy variable such that

$$\delta_i(j) = \begin{cases} 1, & \text{if subject } i \text{ responds to category } j \\ 0, & \text{otherwise} \end{cases} \qquad (41)$$

Then the principle is expressed as follows.

$$\rho = c_{xy}/\sigma_x\sigma_y \rightarrow \max \qquad (42)$$

where

$$c_{xy} = \frac{1}{ln} \sum_{i=1}^{Q} \sum_{j=1}^{R} \delta_i(j) s_i x_i y_i - \left(\frac{1}{ln} \sum_{i=1}^{Q} s_i l_i y_i\right)\left(\frac{1}{ln} \sum_{i=1}^{Q} \sum_{j=1}^{R} \delta_i(j) s_i x_j\right) \qquad (43)$$

$$\sigma^2_x = \frac{1}{ln} \sum_{i=1}^{Q} \sum_{j=1}^{R} \delta_i(j) s_i x^2_j - \left(\frac{1}{ln} \sum_{i=1}^{Q} \sum_{j=1}^{R} \delta_{ij} s_i x_j\right)^2 \qquad (44)$$

$$\sigma_y^2 = \frac{1}{ln} \sum_{i=1}^{Q} s_i l_i y_i^2 - \left(\frac{1}{ln} \sum_{i=1}^{Q} s_i l_i y_i\right)^2 \qquad (45)$$

l_i and $\bar{l}$ denoting the number of categories to which subject i responds and the average over $i = 1, 2, \ldots, Q$.

Using the similar procedure to the formulation based on canonical correlation analysis in the case of the second method, we can easily derive the following eigenvalue problem, when we put

$$\bar{x} = \frac{1}{ln} \sum_{i=1}^{Q} \sum_{j=i}^{R} \delta_i(j) s_i x_j = 0 \qquad (46)$$

for the normalization of location.

$$\mathbf{Hx} = \rho^2 \mathbf{Fx} \qquad (47)$$

where

$$\mathbf{H} = [h_{jk}]$$

where

$$h_{jk} = \sum_{i=1}^{Q} \frac{\delta_i(j)\, \delta_i(k)}{l_i} s_i \qquad (48)$$

where

$$\mathbf{F} = [f_{jk}]$$

where

$$f_{jk} = \delta_{jk} \sum_{i=1}^{Q} s_i \delta_i(k) \tag{49}$$

δ_{jk} indicating Kronecker's delta. Thus the optimal scores $\{x_j\}$ are obtained as the eigenvector corresponding to the largest eigenvalue. The optimal scores $\{y_i\}$ are obtained by

$$y_i = \frac{1}{l_i} \sum_{j=1}^{R} x_j \delta_i(j) \tag{50}$$

If the information is poor by assigning unidimensional scores, we may use multidimensional scores. In such cases the eigenvector corresponding to the eigenvalues smaller than the largest should be used. The principle becomes to maximize $\Pi_i \rho_i$ by assigning multidimensional scores $[x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(t)}]$ and $[y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(t)}]$ under orthogonality conditions,

$$\mathbf{x}^{(j)\prime}\, \mathbf{F}\, \mathbf{x}^{(k)} = 0 \quad \text{for } j \neq k \tag{51}$$

Concerning $\mathbf{y}^{(j)}$, the orthogonality conditions (51) are expressed by

$$\mathbf{y}^{(j)\prime}\, \mathbf{G}\, \mathbf{y}^{(k)} = 0 \quad \text{for } j \neq k \tag{52}$$

where $\mathbf{G} = [\delta_{jk} s_j l_j]$.

As methods similar to or extended from Hayashi's third method there exist the scalogram analysis of Guttman (*2*) and the categorical canonical correlation analysis of Okamoto and Endo (*15, 16*) and so on.

Fourth Method of Quantification (Quantification IV)

Suppose a similarity index e_{ij} is observed between each pair of subjects in a sample of size n, where the similarity index indicates that a pair (i, j) with a large e_{ij} is more similar with each other than a pair (i', j') with a small $e_{i'j'}$. The fourth method of quantification is a method to quantify the subjects on the basis of these similarity indexes, to represent them in an appropriate dimensional Euclidian space and, if it is required, to classify them.

When we assign a numerical score x_i to subject i, the principle for quantification is expressed as follows.

$$Q = -\sum_{i \neq j}\sum e_{ij} (x_i - x_j)^2 \rightarrow \max. \tag{53}$$

$$s_x^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \text{const.} \tag{54}$$

Since Q and s_x^2 are invariant under a shift of origin, we may choose

$$\bar{x} = 0 \tag{55}$$

without any loss of generality. Introducing a Lagrange multiplier λ, the problem (53)-(54) is transformed to

$$L = Q - (s_x^2\text{-const.}) - \sum_{i \neq j}\sum e_{ij} (x_i - x_j)^2 - \lambda\left(\frac{1}{n}\sum_i x_i^2 - \text{const.}\right) \rightarrow \max. \tag{56}$$

Hence we obtain

$$\sum_j h_{ij} x_j = (\lambda/n + \sum_j h_{ij})\, x_i \tag{57}$$

where

$$h_{ij} = h_{ji} = e_{ij} + e_{ji} \tag{58}$$

Since Q does not depend on e_{ij} which is undefined, we may specify as

$$\sum_j h_{ij} = \sum_j (e_{ij} = e_{ji}) = 0 \tag{59}$$

Thus we finally obtain the following eigenvalue problem.

$$\mathbf{H}\, \mathbf{x} = \mu\, \mathbf{x} \tag{60}$$

where

$$\mathbf{H} = [h_{ij}] : n \times n$$
$$\mathbf{x} = [x_i] : n \times 1$$
$$\mu = \lambda/n \tag{61}$$

Obviously, from Eqs. (60) and (61),

$$\lambda = n\, \mathbf{x}'\, \mathbf{H}\, \mathbf{x}/\mathbf{x}'\mathbf{x} = Q/s_x^2 \tag{62}$$

Therefore, the optimal scores are given by the eigenvector $x^{(1)}$ corresponding to the largest eigenvalue μ_1 of Eq. (60), normalized according to Eqs. (54) and (55).

In the case where the classification is not satisfactory with the eigenvector $x^{(1)}$, we may use the eigenvectors corresponding to the second $\sim p$-th largest eigenvalues. The principle becomes to maximize

$$Q = -\sum_{i \neq j}\sum e_{ij} \left\{ \frac{(x_i^{(1)} - x_j^{(1)})^2}{s^2(x^{(1)})} + \frac{(x_i^{(2)} - x_j^{(2)})^2}{s^2(x^{(2)})} + \ldots + \frac{(x_i^{(p)} - x_i^{(p)})^2}{s^2(x^{(p)})} \right\} \tag{63}$$

under the orthogonality conditions

$$\mathrm{cov}\,(x^{(k)}, x^{(l)}) = \frac{1}{n}\sum_i (x_i^{(k)} - \bar{x}^{(k)})\,(x_i^{(l)} - \bar{x}^{(l)}) = 0 \tag{64}$$

by the idea of assigning a multidimensional score $(x_i^{(1)}, \ldots, x_i^{(p)})$ to subject i. The number of dimensions p is determined by the decreasing pattern of $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(n)}$.

According to the above explanation it may be clear that the fourth method of quantification is a kind of multidimensional scaling (MDS), or precisely speaking, a kind of metric MDS in the sense that the result depends on the value of e_{ij} itself instead of the rank order of e_{ij}.

Quantification of Ordered Categories

In the methods of quantification described above, no order relation is supposed among the categories of the qualitative external criterion and/or the qualitative attributes. Even if we have prior information about the order relations in actual data analysis, we sometimes obtain a solution inconsistent with the prior information by applying the ordinary methods of quantification. In such cases it may be appropriate to apply the methods of quantification for ordered categories, which we shall discuss in this section.

Case with Some Order Relations among the Categories of Attributes in the First Method of Quantification

Let us introduce inequality constraints corresponding to order relations among the categories of attribute items. Now that the first method of quantification is mathematically equivalent to the multiple regression analysis on dummy variables, we may formulate the problem of quantification for ordered categories as the problem of regression with some inequality constraints. Then we must solve the optimization problem (4) under some inequality constraints such as, for example,

$$s_{j1} \geq s_{j2} \geq \ldots \geq s_{jc(j)} \tag{65}$$

Since the constraints are generally linear, we can reformulate the problem as in the case of the second method of quantification and solve it iteratively but efficiently by using Wolfe's reduced gradient procedure. Furthermore, if we make use of the property that the mean square error is quadratic with respect to $\{s_{jk}\}$, we can solve the problem more efficiently by the quadratic programming technique.

Case with Some Order Relations among the Categories of the External Criterion in the Second Method of Quantification

Although the categories of the external criterion are defined as nominal in the ordinary method, we sometimes meet the situations with ordinal external criteria. For example, in medical research we meet situations in which the severity rating, improvement rating, or sometimes the movement of severity rating should be chosen as the external criterion and we wish to analyze the effects of factors on it.

According to the formulation based on canonical correlation analysis, the optimal score vector is obtained as the solution of Eq. (66)

$$\rho^2 = \tilde{\mathbf{t}}'\,\tilde{\mathbf{S}}_{12}\,\tilde{\mathbf{S}}_{22}^{-1}\,\tilde{\mathbf{S}}_{21}\,\tilde{\mathbf{t}}/\tilde{\mathbf{t}}'\tilde{\mathbf{S}}_{11}\tilde{\mathbf{t}} \rightarrow \max. \tag{66}$$

Thus the problem becomes to maximize the nonlinear objective function under an arbitrary set of order restrictions such that

$$t_j \geq t_{j'} \qquad (j, j') \in S \tag{67}$$

where S denotes a set of pairs of subscripts corresponding to the categories ordered theoretically.

The problem of quantification under order restrictions was studied by Bradley, Katti, and Coons (*17*), Nishisato and Arri (*18*), Tanaka and Asano (*19*), Tanaka, Asano, and Kodake (*20*), and Tanaka (*21*), among others. Bradley et al. (*17*) solved the case of complete order restrictions. Nishisato and Arri (*18*) extended it to the case of a special type of partial order restrictions, and we solved the case of arbitrary order restrictions generally (*19-21*).

As shown previously (*19, 21*), the optimization problem [Eqs. (66), (67)] can be always transformed to the optimization problem under constraints of nonnegativeness and linear equalities such that

$$\rho^2 = \mathbf{z}'\,\mathbf{C}\,\mathbf{z}/\,\mathbf{z}'\,\mathbf{D}\,\mathbf{z} \rightarrow \max. \tag{68}$$

subject to

$$z_{(k)} = [z_{(k)1}, z_{(k)2}, \ldots, z_{(k)c_k}]' \geq 0 \qquad k = 1, 2, \ldots, m \tag{69}$$

$$\mathbf{a}'_{(kj)}\mathbf{z}_{(k)} = 0 \qquad j = 1, 2, \ldots, c_k - r_k + 1 \qquad k = 1, 2, \ldots, m \tag{70}$$

where $z' = [z'_{(0)}, z'_{(1)}, \ldots, z'_{(m)}]$. After this transformation the numerical solution can be obtained efficiently by applying Wolfe's reduced gradient method. As a numerical example, Table 5, which shows the data for a five-treatment experiment with a five-point scoring scale, is taken from the study of Bradley, Katti, and Coons (*17*). Let us suppose the order restrictions $t_1 \geq \{t_2, t_3\} \geq t_4 \geq t_5$ artificially and

Table 5. Numerical example.[a]

Treatment	Response 1 (t_1)	2 (t_2)	3 (t_3)	4 (t_4)	5 (t_5)	Total
1	9	5	9	13	4	40
2	7	3	10	20	4	44
3	14	13	6	7	0	40
4	11	15	3	5	8	42
5	0	2	10	30	2	44
Total	41	38	38	75	18	210

[a]Data of Bradley et al. (*17*).

apply the generalized method, where $a \gtrless \{b, c\}$ denotes $a \gtrless b$ and $a \gtrless c$. These restrictions are expressed by Figure 1.

Then the problem becomes

$$\mathbf{Q} = \mathbf{z}'\,\mathbf{C}\,\mathbf{z}/\mathbf{z}'\,\mathbf{D}\,\mathbf{z} \rightarrow \max. \tag{71}$$

subject to

$$\mathbf{z} = [z_1, z_2, \ldots, z_5] \geq 0 \tag{72}$$

$$z_1 - z_2 + z_3 - z_4 = 0 \tag{73}$$

where

$$\mathbf{C} = \begin{bmatrix} 0.310250 & 0.252340 & 0.336210 & 0.394120 & 0.162770 \\ 0.252340 & 2.099860 & 2.888190 & 1.040670 & -0.280090 \\ 0.336210 & 2.888190 & 3.988550 & 1.436570 & -0.300860 \\ 0.394120 & 1.040670 & 1.436570 & 0.790020 & 0.141990 \\ 0.162770 & -0.280090 & -0.300860 & 0.141990 & 0.835500 \end{bmatrix} \tag{74}$$

$$\mathbf{D} = \begin{bmatrix} 12.998770 & 3.498790 & -0.210730 & 9.289268 & 1.757150 \\ 3.498790 & 12.998779 & 9.289268 & -0.210730 & 1.757150 \\ -0.210730 & 9.289268 & 17.703537 & 8.203548 & 5.014279 \\ 9.289268 & -0.210730 & 8.203548 & 17.703537 & 5.014279 \\ 1.757150 & 1.757150 & 5.014279 & 5.014279 & 16.457138 \end{bmatrix} \tag{75}$$

The equality restriction (73) corresponds to the circuit t_1—t_2—t_4—t_3—t_1 in Figure 1.

The application of the reduced gradient method to the problem of Eqs. (71)-(73) yields the result shown in Table 6. Normalizing so as to satisfy $t_1 = 1.0$ and $t_5 = 0.0$, the optimal scores are given as

$$t = [1.0000 \quad 1.0000 \quad 0.1435 \quad 0.0000 \quad 0.0000]'$$

Case with Some Order Relations among the Categories of the Attributes in the Second Method of Quantification

Suppose there exist some order restrictions such that

$$s_{j1} \geq s_{j2} \geq \ldots \geq s_{jc(j)}$$

or

$$s_{j1} \leq s_{j2} \leq \ldots \leq s_{jc(j)} \qquad j \in J \tag{76}$$

where J denotes a set of subscripts for the items with ordered categories. Then the problem becomes to maximize the nonlinear objective function (16) under the inequality restrictions (76) and can be solved generally by the procedure described above.

In the discriminant analysis using the quantified qualitative variables, we sometimes meet situations where each of the order restrictions may be ascending or descending, i.e.,

$$s_{j1} \geq s_{j2} \geq \ldots \geq s_{jc(j)}$$

or

$$s_{j1} \leq s_{j2} \leq \ldots \leq s_{jc(j)} j \in J \tag{77}$$

This type of quantification was discussed by Tanaka, Asano, and Kubota (22).

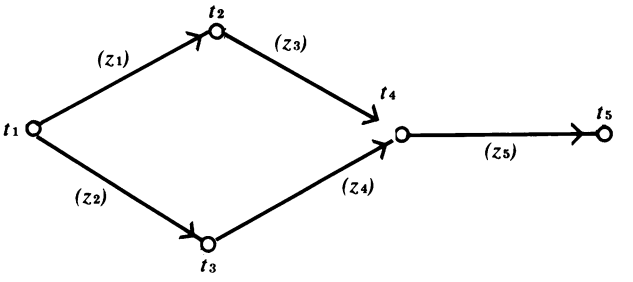

FIGURE 1. Order restrictions.

Case where there exist some order relations in the third method of quantification

Suppose there exist some order restrictions such that

$$y_i \geq y_{i'} \qquad (i, i') \in T$$

or

$$x_j \geq x_{j'} \qquad (j, j') \in S \tag{78}$$

where T and S denote sets of subscripts for the pairs of subjects and of categories with order relations, respectively. In fact we sometimes meet the cases where the attribute categories are divided into several items and the categories in each item are ordered theoretically. We also meet the cases where some order relations are present among the subjects. These problems can be solved by the second procedure above.

Kruskal (*23*) treated a situation analogous to the latter cases and considered to rescale numerical measurements so that on ANOVA model fits as well as possible. He applied an algorithm which uses isotonic regression like in his nonmetric multidimensional scaling. Recently, de Leeuw, Young, and

Table 6. Solution under $t_1 \geq \{t_2, t_3\} \geq t_4 \geq t_5$.[a]

Cycle	z_1	z_2	z_3	z_4	z_5	$Q(z)$
0	*1.00000	1.00000	1.00000	1.00000	1.00000	0.1224780
1	*$0.0^4$757	1.10058	1.70579	0.60522	0.58841	0.1978295
2	0.06542	*1.10058	1.70580	0.60521	0.58839	0.1978315
3	$0.0^8$964	*1.85925	2.03080	0.26120	$0.0^5$930	0.2423639
4	$0.0^5$140	*1.75611	2.03277	0.27666	$0.0^5$930	0.2435769
5	$0.0^5$140	*1.74103	2.03097	0.28994	$0.0^5$930	0.2435847
6	$0.0^5$140	*1.74063	2.03092	0.29029	$0.0^5$930	0.2435847
7	$0.0^5$140	*1.73886	2.03070	0.29184	$0.0^5$930	0.2435848
8	$0.0^5$140	*1.73990	2.03084	0.29094	$0.0^5$930	0.2435847
9	$0.0^5$140	1.73944	2.03078	0.29134	$0.0^5$930	0.2435848

[a]Asterisk (*) indicates that it is selected as a basic variable in each cycle.

Takane (*24*) generalized Kruskal's method and proposed the alternating least squares algorithm. Comparing with these two methods, our method has the following advantages and disadvantages (*25*).

Advantages. It is applicable to generalized criteria for optimal scaling such as CS-1-5, CM-1-7 proposed previously (*12-14*). It is also applicable to the cases with arbitrary partial order relations. The rapidness of convergence depends only on the number of ordered categories, say p. Thus it can be efficiently used when p is small.

Disadvantages. It does not converge rapidly when p is large.

Statistical Considerations

Few statistical considerations of quantification had been studied until comparatively lately. Okamoto and Endo (*16*) investigated the asymptotic distribution of the sample optimal scores for their categorical canonical correlation analysis, which was proposed as a generalization for third method of quantification. Tanaka and Asano (*12, 13*) and Tanaka (*14*) studied the statistical inference of factor-response relationships as well as the asymptotic distribution of the optimal scores based on their CS-1-5 and CM-1-7 criteria, which were proposed as generalizations for the second method of quantification. Although the probabilistic models introduced should be evaluated if they fit to the actual data, the methods will be useful when the sample sizes are large enough to be analyzed by asymptotic theories.

Consider the case of the second method of quantification, where there exist a response and several factors, and the response is chosen as the external criterion. For such cases the probabilistic model shown in Figure 2 has been proposed (*12-14*).

As shown in the preceding section, the optimal scores are determined by an eigenvalue problem such that

$$(\mathbf{A} - \lambda \mathbf{B})\,\mathbf{t} = 0$$

By means of the δ-method, small deviations of the eigenvalues and vectors can be asymptotically approximated by linear equations of the small deviations of the matrices **A** and **B**, under the assumption that the eigenvalues are all distinct. Furthermore, small deviations of the elements of the matrices **A** and **B** can be expressed by the Taylor expansions of the multinomial proportions on the basis of the above probabilistic model. Thus, as a result, the small deviations of the eigenvalues and vectors (optimal scores) are asymptotically approximated by the functions of the small deviations of the multinomial proportions. From this, the asymptotic normality of the sample optimal scores are derived.

Computer Programs

The use of electronic computers is indispensable in applying the methods of quantification, because the calculations are complex and ordinarily a comparatively large amount of data are analyzed by these methods. One of the reasons that Hayashi's four methods are widely applied in Japan may be that the program packages are available to the data analysis. They are, for example, Component Analysis 1-4 (IBM-Japan), Quantas 1-4 (FACOM), Firms III (NEAC), Quantification 1-4 (Dentsu MARK III),

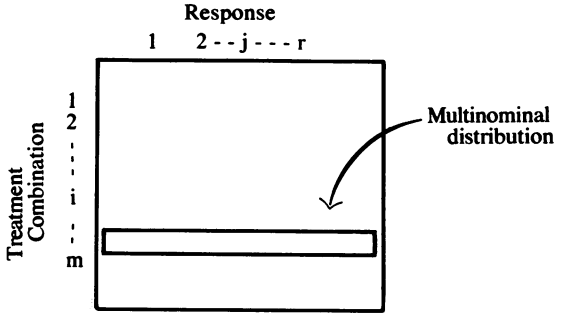

FIGURE 2. Probabilistic model.

Hayasi 1 ~ 4 (SPSS-Japanese Version), and so on. Furthermore, in the NISAN system (*26*), now being developed by a group of Japanese statisticians, the varieties of methods including those for ordered categories and based on the asymptotic theories will be available for the convenience of senior statisticians.

It may be obvious from the derivations in the previous sections that the methods of quantification, especially from the first to third methods, are mathematically equivalent to regression analysis, canonical analysis, and canonical correlation analysis applied to dummy variables corresponding to categorical data. Therefore, if we carefully use the programs, we can apply the methods of quantification to data analysis by means of the programs for ordinary multivariate analyses.

REFERENCES

1. Fisher, R. A. Statistical Methods for Research Workers. 10th ed. Oliver and Boyd, London, 1946.
2. Guttman, L. The quantification of a class of attributes: a theory and method of scaling construction. In: The Prediction of Personal Adjustment, P. Horst, Ed., Social Science Research Council, New York, 1941.
3. Hayashi, C. On the quantification of qualitative data from the mathematico-statistical point of view. Ann. Inst. Statist. Math. 2: 35 (1950).
4. Hayashi, C. On the prediction of phenomena from qualitative data on the quantification of qualitative data from the mathematico-statistical point of view. Ann. Inst. Statist. Math. 3: 69 (1952).
5. Hayashi, C. Multidimensional quantification — with the applications to the analysis of social phenomena. Ann. Inst. Statist. Math. 5: 121 (1954).
6. Hayashi, C. Sample survey and theory of quantification. Bull. Inst. Statist. Inst. 38: 505 (1961).
7. Hayashi, C., and Murayama, T. Planning and Practice of Marketing Research. Nikkan Kogyo Publishing Co., Tokyo, 1964.
8. Hayashi, C., Higuchi, I., and Komazawa, T. Data Processing and Statistical Mathematics. Sangyo Tosho Publishing Co., Tokyo, 1970.
9. Hayashi, C. Methods of Quantification. Toyo Keizai Shinpo Publishing Co., Tokyo, 1974.
10. Rao, C. R. Linear Statistical Inference and Its Applications. John Wiley, New York, 1965.
11. Johnson, P. O. The quantification of qualitative data in discriminant analysis. J. Am. Statist. Assoc. 45: 65 (1950).
12. Tanaka, Y., and Asano, C. Some generalized methods of optimal scaling and their asymptotic theories: the case of single response-multiple factors. Research Institute of Fundamental Information Science, Kyushu Univ., Research Report No. 77 (1978).
13. Tanaka, Y., and Asano, C. Some generalized methods of optimal scaling and their asymptotic theories. Proceedings of International Conference on Quality Control, Tokyo, Oct. 17-20, 1978.
14. Tanaka, Y. Some generalized methods of optimal scaling and their asymptotic theories: The case of multiple responses-multiple factors. Ann. Inst. Statist. Math. 30: 249 (1978).
15. Okamoto, M., and Endo, H. Basic properties of categorical canonical correlation analysis. J. Japan. Statist. Soc. 4: 15 (1973).
16. Okamoto, M., and Endo, H. An asymptotic theory of categorical canonical correlation analysis. J. Japan. Statist. Soc. 5: 1 (1974).
17. Bradley, R. A., Katti, S. K., and Coons, I. J. Optimal scaling for ordered categories. Psychometrika 27: 355 (1962).
18. Nishisato, S., and Arri, P. S. Nonlinear programming approach to optimal scaling of partially ordered categories. Psychometrika 40: 525 (1975).
19. Tanaka, Y., and Asano, C. A generalized method of optimal scaling for partially ordered categories. Proceedings of the Third Symposium on Computational Statistics, Leiden, the Netherlands, Aug. 21-25, 1978.
20. Tanaka, Y., Asano, C., and Kodake, K. Application of nonlinear programming techniques to the problem of optimal scaling for ordered categories. Res. Inst. Fund. Inf. Sci., Kyushu Univ., Res. Rept. No. 85 (1978).
21. Tanaka, Y. Optimal scaling for arbitrarily ordered categories. Ann. Inst. Statist. Math., in press.
22. Tanaka, Y., Asano, C. and Kubota, N. A generalized method of optimal scaling for multiple responses with ordered categories. Res. Inst. Fund. Inf. Sci., Kyushu Univ., Res. Rept. No. 86 (1978).
23. Kruskal, J. B. Analysis of factorial experiments by estimating monotone transformation of the data. J. Roy. Statist. Soc. B27: 251 (1965).
24. de Leeuw, J., Young, F. W. and Takene, Y. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. Psychometrika 41: 471 (1976).
25. Tanaka, Y., Kodake, K., Kubota, N. and Asano, C. Optimal scaling for ordinal measurements in NISAN system. Res. Inst. Fund. Inf. Sci., Kyushu Univ., Res. Rept. No. 89 (1979).
26. Asano, C., Wakimoto, K., Shohoji, T., Jojima, K., Goto, M., Tanaka, Y., Tarumi, T. and Ohsumi, N. The statistical principle and methodology in Nisan system. Proceedings of the Third Symposium on Computational Statistics, Leiden, the Netherlands, Aug. 21-25, 1978.